Australian Bureau of Statistics

1352.0.55.151 - Research Paper: A New Analytical Platform to Explore Linked Data (Methodology Advisory Committee), June 2015

Latest ISSUE Released at 11:30 AM (CANBERRA TIME) 10/08/2015 First Issue

Summary

Executive Summary

EXECUTIVE SUMMARY

The ABS is exploring semantic web techniques as one possible solution to overcome some of the challenges in the information age. Some examples include:
· increasing user demand for more timely and sophisticated statistical products;
· exploring efficient data management systems for data integration which are flexible and cost effective;
· maintaining conceptual coherence from many data sources; and
· researching new methods for the analysis and visualisation of networks in Big Data to unlock new insights.

This paper describes a prototype Graphically Linked Information Discovery Environment (GLIDE) created using semantic web techniques to better manage statistical information. The Semantic Web framework provides an alternative approach to data representation, linking and retrieval that can unlock the full potential of interconnected and multi-dimensional datasets. Instead of organising datasets in a structured row-column tabular form, the Semantic Web approach models information in the form of a network of entities and relationships. The relationships are given strong computable semantics by precisely specifying their logical properties in a machine-interpretable format. This allows computers to understand these relationships to easily explore multiple data dimensions to identify interesting data patterns and analyse the network structure of data.

This paper demonstrates one analytical application of the GLIDE by using it to derive network statistics and create models to distinguish true firm deaths from spurious ones. The ABS has an established process for identifying firm exits, but is not able to distinguish the type of exit – whether it is due to restructuring, merger/takeover or a genuine death.

The analytical results have shown that it is important to account for spurious death for statistical production. This is because failure to account for spurious firm deaths can result in continuing enterprises being incorrectly classified as firm deaths and as a result it can affect the statistical quality from the perspectives of survey frame and accuracy of the statistics.

This paper considers both multilevel and Bayesian Networks (BNs) models. Our approach applies the BNs method within a statistical framework. We have shown that BNs can handle observations with missing variables in the test data. This paper does not intend to compare both methods on the prediction outcomes. It clearly shows that it is important to incorporate network information for modelling purposes. This leads to the prediction outcomes improved substantially for both models, reaching a 95% accuracy rate.

We conclude that the semantic web is a useful approach for statistical purposes and that network analysis can be used to effectively distinguish true and spurious firm deaths.

About this Release

The advancement of technology, new methods and emerging data sources have presented both opportunities and challenges to the ABS. While Big Data provides new business opportunities for statistical production, there remain some challenges the ABS needs to overcome. The ABS is exploring semantic web techniques as one possible solution to overcome some of these challenges. This paper describes a prototype Graphically Linked Information Discovery Environment (GLIDE) created using semantic web techniques to better manage statistical information. This paper demonstrates one analytical application of the GLIDE by using it to derive network statistics and create models to distinguish true firm deaths from spurious ones. The ABS has an established process for identifying firm exits, but is not able to distinguish the type of exit – whether it is due to restructuring, merger/ takeover or a genuine death. This paper uses multilevel and Bayesian Network models to distinguish true and spurious firm deaths by incorporating network statistics. It is important to account for spurious deaths for statistical production to ensure data quality. The model results also perform much better after incorporating network statistics. We conclude that semantic web is a useful approach for statistical purposes and that network analysis can be used to effectively distinguish true and spurious firm deaths.